

## **Adding typology to lexicostatistics: a combined approach to language classification**

Dik Bakker, Cecil H. Brown, Pamela Brown, Dmitry Egorov, Anthony Grant, Eric W. Holman, Robert Mailhammer, André Müller, Viveka Velupillai, Søren Wichmann<sup>1</sup>

### **0. Introduction**

The ASJP project aims at developing a general method for the automatic reconstruction of relationships between languages. Such reconstructions may be instrumental in the critical assessment and refinement of existing genetic and areal classifications. Likewise, irregularities in phylogenies may be detected, while newly described and as yet unclassified languages may be classified. Furthermore, methods for dating languages and language groups may be tested and enhanced experimentally. Finally, the method may be helpful in distinguishing between cognates and borrowings. All these derived goals are pursued by members of the project. The basis of the reconstruction process is a matrix with distances between individual languages, dialects and protolanguages (in short: between languages). In its original conception these distances were based exclusively on the number of cognates for pairs of languages that may be found in basic word lists such as the one proposed by Swadesh (1955). In section 1 we will give a brief sketch of this purely lexical strategy. Although it was shown in Brown et al (2008) that classifications based on this approach come fairly close to well-established classifications presented in the literature, later experiments in Holman et al (2008) have made it clear that an exclusively lexicostatistical method may be made more accurate by the addition of typological data. The database used for the World Atlas of Language Structures (henceforth WALs; Haspelmath et al. 2005), to which several of the ASJP members contributed, is the obvious source for such data. For this exercise, we selected what we think are the most stable features from WALs. In section 2 we will summarize the criteria for this selection, which are presented in full by Wichmann and Holman (n. d.). In section 3 we will explore the reliability of a classification purely based on a subset of the most stable WALs features. Like the purely lexical classification, this turns out to give quite reliable results. However, there is room for improvement. In section 4 we will have

a look at the results of an approach, also described in more detail in Holman et al (2008), that integrates both types of data. Section 5 draws some conclusions.

### **1. The lexicostatistical approach**

Ever since its inception in 1955, the list of basic words compiled by Morris Swadesh has been an instrument for establishing genetic relations between languages. In its attempt to develop a general method to find relationships between languages, ASJP has taken the original 100-item list as a point of departure. However, our method differs from the original lexicostatistical approach in two fundamental ways. First, the comparison between word forms is done by a computer program on the basis of Levenshtein's (1966) algorithm, as in the recent work of Serva and Petroni (2008), resulting in a distance matrix between individual languages. And second, graphic branching structures illustrating language relatedness (rooted or unrooted family trees) are generated from this matrix by the way of standard software and algorithms originally developed for the use of biologists in studying phylogenetic relationships (cf. Huson 1998). To accommodate wordlists originally published in a variety of more or less simplified orthographies, a special alphabet, called ASJPcode, was devised which makes use of the QWERTY keyboard symbols only. It contains just 34 consonant symbols and 7 symbols for vowels. These symbols are used for phonological segments defined by the most common points and manners of articulation. Rarer segments are represented by the symbol they most closely resemble in terms of point and manner of articulation. See Brown et al (2008) for details.

Unlike most other approaches to automatic language classification, such as those described by Oswalt (1971), Atkinson et al. (2005), and Nakhleh et al. (2005), the present method automates both the judgments of cognacy and the subsequent inference of phylogeny. We can therefore apply the same objective criteria to classify an unusually large sample of languages. In fact, the project strives for completeness rather than for some representative sample of the languages of the world. This facilitates the large scale statistical study of overlaps in lexicons between languages and may reveal previously unknown phylogenetic relationships. To date, we have collected and transcribed a basic word set for almost 2,500 languages of the world. The around three million language

pairs in this database are compared by means of the Levenshtein Distance measure (LD; see Levenshtein 1966), which we now use instead of the distance measure described by Brown et al. (2008) because it produces better agreement with published classifications. For any pair of words represented in ASJPCODE, LD is defined as the minimum total number of additions, deletions, and substitutions of symbols necessary to transform one word into the other. For any pair of languages L1 and L2, first the LD values are established for each of the N Swadesh words that L1 and L2 share (virtually always the full set that we consider). These LD values are then normalized by dividing each LD by its theoretical maximum, giving the normalized LD (LDN). Finally, since lexical similarity may be influenced by chance resemblances, such as an overlap in the phoneme inventories or shared phonotactic preferences for the two languages involved, we correct each LDN by dividing it by the mean LDN of all  $N(N-1)/2$  pairings of words with different meanings, giving the LDND value for each of the N meaning pairs. The LDND value for the language pair L1 – L2, i.e. their Levenshtein distance, is defined as the mean of the LDND values for the individual word pairs.

Earlier experiments in Holman et al (2008) on 245 languages have shown that the 100-item Swadesh list may be reduced to a much shorter one, without loss and even with a gain in classificatory reliability. The subset we selected contains the 40 most stable elements from the original list. They are given in Table 1 below, which is extracted from the appendix of Holman et al (2008).

I	Leaf	Knee	Star
You	Skin	Hand	Water
We	Blood	Breast	Stone
One	Bone	Liver	Fire
Two	Horn	Drink	Path
Person	Ear	See	Mountain
Fish	Eye	Hear	Night
Dog	Nose	Die	Full
Louse	Tooth	Come	New
Tree	Tongue	Sun	Name

Table 1. The 40 most stable Swadesh items

The strategy employed to establish this most stable subset is similar to the one used by Wichmann and Holman (n. d.) to establish the relative stability of the typological features of the WALS database, which will be summarized in the next section. Suffice it for now that the basic idea is that we compare the outcome of our classifications on the basis of different subsets of the full Swadesh list with the families and genera as established by Dryer (2005) as used by WALS, and the genetic classification of the *Ethnologue* (Gordon 2005). If we take these two classifications as a point of departure, then iterative comparisons led us to the specific subset of 40 items in table 1. It makes better predictions than any smaller subset, and at least as good predictions as any larger subset.

Genetic reconstructions based on this purely lexicostatistical method appear to be very promising. However, languages also inherit grammatical features. In section 3 we will see how an approach fares that makes use of grammatical data, stemming from the WALS database. First, however, we will see how we may single out the most stable features for this purpose.

## **2. Selecting the most stable WALS features**

Although the WALS project has a purely descriptive goal, and does in no way seek to contribute directly to genetic reconstruction, we think that the wide range and the quality of this database make it a natural first choice for this exercise. However, certain aspects of the WALS data matrix make the operation somewhat problematic. WALS contains 139 features for a total of 2,558 languages.<sup>2</sup> Quite a few of these are mutually dependent in the sense that their values partially or completely overlap. In the case there is an overlap between the data stemming from two different WALS authors, their definitions for the same linguistic category are not necessarily the same. Furthermore, the WALS data matrix is sparse: only around 16% of the cells have an actual value. And not only is there a vast difference between the number of languages for the respective WALS features, ranging between 116 and 1343. There is also no sizeable subset of languages for which a value for all features is available. The best represented subset of 200 languages

has a coverage of around 75% of the features. Nor does the full range of WALS languages adhere to any known type of genetic or areal stratification. There are languages from over 200 families, including isolates. Overall, 458 genera in terms of Dryer (1992) may be distinguished. But although the subsets selected by the individual authors may (but not necessarily do) form a genetically or areally representative sample, the aggregated database certainly does not. As was discussed in Bakker (2008), the largest subset among the WALS languages that falls within the limits of the sampling technique proposed by Rijkhoff & Bakker (1998) is 359. Extension to 360 or more would require a language from a genetic (sub)grouping that is not represented in WALS. And most of the languages in the 359 sample have no value for the majority of the variables. Therefore it is obvious that some kind of selection from the database is necessary for our exercise, both in terms of languages and features. Since we want to use these data to help us establish relationships between languages, we would like to select those features which are the least subject to change, either diachronically or through language contact. In other words: we want to select the most stable features among the WALS set.

Sapir (1970 [1921]: 172) was probably the first to observe that after the split of a language into two dialects, certain features would drift apart more rapidly than others. The latter he assumes to be more fundamental than the former for the characterization of the respective languages. The term ‘stability’ is introduced as such in Greenberg (1978), where it is explicitly linked to frequency, universals of language, and genetic classification. Nichols (1995) might be the first to develop a set of metrics for the measurement of feature stability. One is particularly relevant for the type of data in WALS. It is based on the number of pairs of languages in a genetic or areal group that have the same value for some feature in relation to the number of pairs that have a value for the feature at all.<sup>3</sup> If, for some group, this figure is at least one standard deviation above the mean for all groups, the corresponding value is considered to be characteristic for the group in question, and presumably stable.

The metric proposed by Wichmann and Holman (n. d.) may be seen as a refinement of the one proposed by Nichols. The genetic classification taken as a point of departure consists of Dryer’s genera as well as the family classification used in the WALS database. As in Nichols (1995), the number of language pairs within each genus

G that have the same value for a feature F is counted. This number is divided by the total number  $T_F$  of language pairs in G which have any value for F. This proportion of languages with the same value for F in G is then weighted by the square root of  $T_F$ . This weighting largely neutralizes the disproportionate effect of differences in absolute numbers of languages between groups. These weighted proportions are averaged over all genera and called R, for similarly featured Related languages. R gives an impression of the amount of similarity for some feature across genetic or areal groups. However, R should be controlled for the probability that languages share their value for F by chance. Therefore, U is calculated as the proportion of all pairs of Unrelated languages, stemming from different families, which have the same value for F. This could be seen as the baseline for equality on feature F. U is subtracted from R to give a more realistic estimate of the consistency of the value for F within groups. Finally, the stability factor  $S_F$  for F is defined by dividing the difference between R and U by the maximum possible distance from the baseline, as in (1) below. This normalization allows for direct comparison between the respective S values for the different features in the database, and is a common operation for similarity coefficients (cf. Albatineh et al. 2006).

$$(1) \quad S_F = (R - U) / (1 - U)$$

$S_F$  has maximum value 1.0 in case all languages in the groups have the same value for F, and minimum expected value 0.0 in case unrelated languages are as similar on the relevant feature value as related ones.<sup>4</sup>

A further complication is introduced by language contact and change as a result of it. The value of R is based on language pairs within the same genus. These are typically geographically close to each other. The language pairs that determine the value of U stem from different families, and may be either close or distant. To compensate for this potential factor, only those language pairs are included which are located less than 5,000 km from each other, according to the coordinates provided by WALS. This is a compromise between 20,000 km, which would include all language pairs, but would maximize the difference for the language contact factor, and a value lower than 5,000 km, which would decrease the difference of the contact factor but would make the

empirical basis for the results increasingly smaller. Holman et al. (2007: 397) have shown that beyond a 5,000-6,000 km range the effects of language contact on typological similarities among languages are negligible.

Table 2 gives an impression of the results of this operation for the WALS database. We give the five most stable features (top) and the five most unstable ones (bottom), extracted from appendix 1 of Wichmann and Holman (n. d.).

<b>WALS Variable</b>	<b>Description</b>	<b>Stability</b>
31	Sex-based and Non-sex-based Gender Systems	0.81
118	Predicative Adjectives	0.74
30	Number of Genders	0.73
119	Nominal and Locational Predication	0.71
29	Syncrretism in Verbal Person/Number Marking	0.71
...		
128	Utterance Complement Clauses	0.07
115	Negative Indefinite Pronouns and Predicate Negation	0.07
59	Possessive Classification	0.01
135	Red and Yellow	-0.07
58	Obligatory Possessive Inflection	-0.25

Table 2. The most stable and unstable WALS features

We are now in the position to see whether typological features of grammar such as the ones in WALS are as useful for measuring distances between languages as are lexical items from the Swadesh list.

### **3. Using WALS features for establishing language relations**

For this exercise we proceeded as follows. We selected the 500 languages in the WALS database with the highest numbers of attested values. Starting out with the 130 most stable features, we calculated the distance between all 124,750 pairs of languages. The distance between two languages was defined as the proportion of the relevant features for which both languages had a different value. On the basis of the resulting distance matrix we assessed its classificatory performance with respect to the combined WALS genetic classifications, the one for genera and the one at the family level. We then excluded the five least stable features from the process, recalculated the distance matrix and compared

the results again with this classification. This was done 24 times iteratively until we were left with the ten most stable features. Figure 1 shows the results of these comparisons. The horizontal axis represents the number of most stable features taken into consideration. The vertical axis gives the correlation with the WALS classification.

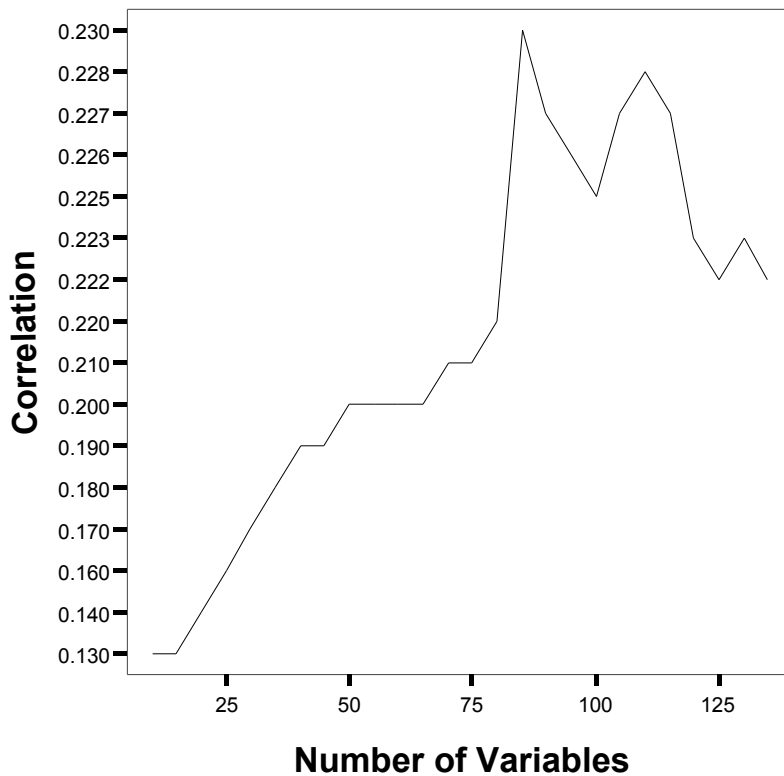


Figure 1. Correlation between WALS features and WALS classification

It is clear that, now going from left to right and adding more, be it less stable features the correlations get better. However, this seems to peak around the 85 most stable features. Adding more features to this does not improve the correlation. This suggests that our ideal sample would contain all languages with a value for the 85 most stable WALS features.

A final observation we want to make here concerns the genetic classification we used. WALS only provides a distinction at two levels: genus and family. This provides

one with a very restricted number of different genetic distances between languages: 0 (same genus), 1 (same family but different genus) and 2 (different family). Comparisons we made with the Ethnologue classification (Gordon 2005), with its more refined subdivisions within families, show that one gets substantially higher correlations in that case, typically between 0.6 and 0.8 (cf. figure 4 in Holman et al. 2008).<sup>5</sup>

Nevertheless, even our best sample, based on the 85 most stable WALS features, performs more or less equally well in terms of classifying properties as the sample based on the 40 most stable Swadesh items, described in section 1. Were classification to be the sole purpose of the sample, then the latter would be the one to be preferred, since it is much easier to construct in terms of data collection. This however leaves open the possibility that a combined sample would do better than any of the two separately. To explore this is the objective of the next section.

#### **4. Merging lexical and grammatical data**

In section 1 we established a subset of the Swadesh list which was at the basis of a distance matrix for the 2400-plus languages in our database. Although from this set we inferred classifications that come very close to the ones found in the literature there is some room for improvement. In the end, we would want to be sure that differences from existing classifications pointed out by the ASJP method are reliable enough to lead to further exploration and eventually even to a reassessment of those classifications and the allocation of so far unclassified languages. Refining the sample with the most stable features from the WALS seems like a promising strategy. Of course ideally, for this purpose we would need a complete overlap in data between the two databases. Given the incompleteness of the WALS database, this is problematic. We adopted the following strategy. We selected the same subset of 500 languages from the WALS database that we used above to establish the correlations for the WALS features. For 355 of these languages we currently have the lexical data available as well.

For this subset we calculated the distances between all language pairs in terms of the genetic algorithm of the SplitsTree program (cf. Huson 1998), taking into consideration the 85 most stable variables in WALS as discussed in section 3. This gave us 62,835 distance pairs. These were compared to the corresponding LDND values for

the same pairs. The Pearson correlation between these two sets of distances turned out to be highly significant (0.063, significant at the 0.001 level). This conclusion got extra support from a Mantel test. A total of 10,000 randomizations of the original languages gave Pearson correlation values between 0.043 and -0.050 for each of them, and a mean of 0.0. The original, non-randomized correlation is more than 6 standard deviations away from the mean.

The observed correlation is significantly above 0 but far below 1. From this we may conclude that the two approaches, one based on the 40 most stable Swadesh items and one based on the 85 most stable WALS variables, are consistent but not redundant. We are now in the position to see to what extent a mix of both data sets does even better than either one separately. In order to determine what the best mix would be we weighted the contribution from both measures of distance for a language pair to an overall distance measure. We started out with 0% lexical distance and 100% typological distance, then 1% LDND distance plus 99% WALS, etcetera, ending with a completely lexical measurement. The results of the correlations with the genetic classification of the WALS are given in figure 2 below.

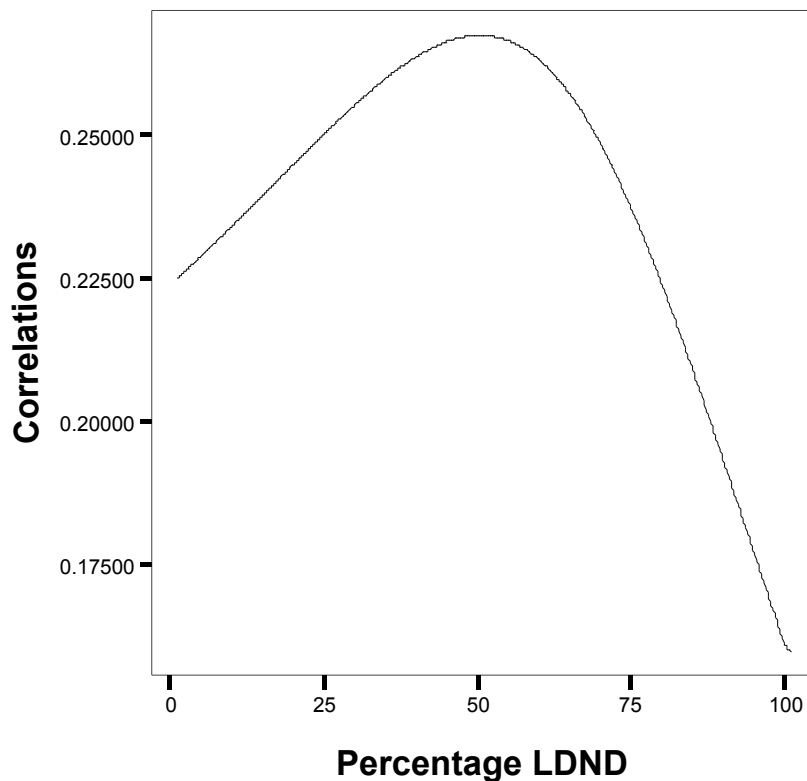


Figure 2. Combining lexical and grammatical data

In figure 2, on the far left we find the results based purely on the WALS distances, a correlation of around 0.22.<sup>6</sup> To the far right we find the results for the purely Levenshtein distances, a correlation of around 0.16. However, a slightly better correlation is reached for a combination of the two factors. The maximum of around 0.27 is reached with a contribution of 50% based on lexical elements and 50% based on typological features.

As earlier on, for the correlation between the LDND and WALS data sets, we have tested these correlations by means of the Mantel test, this time applying 1,000 randomizations.<sup>7</sup> In table 3 below we give some statistics resulting from this experiment.

Percentage LDND vs. WALS	Real Correlation	Standard Deviations From Mean	Mean	Standard Deviation	Highest Random Correlation
0 vs. 100%	0.22492	25	0.00514	0.00905	0.02326

25 vs. 75%	0.25121	26	0.00593	0.00973	0.02751
50 vs. 50%	0.26729	27	0.00672	0.00999	0.02389
75 vs. 25%	0.23465	26	0.00652	0.00899	0.02528
100 vs. 0%	0.16117	23	0.00501	0.00712	0.02294

Table 3. Statistics for some distributions of lexical and typological data

From these figures it is not only clear that the correlations increase to the extent that the weights for the two data sets are more equivalent (columns 1 and 2). We can also see that the correlation for the 50/50% distribution is more strongly positive in terms of number of standard deviations from the mean than the other distributions (columns 3 through 5). In the rightmost column one finds the highest correlation for any of the randomly generated language pairs, which is typically around 10 times closer to the mean than the real correlation.

## 5. Conclusion

In this contribution we have tried to show that distance matrices based on either basic word lists or sets of typological features may provide us with classifications that very closely match some of the most well-known global genetic classifications in existence to date. A combined approach, based for around 50% on lexical data and 50% on typological features works even better. An optimal result is reached when the 40 most stable lexical items from the Swadesh list are combined with the 85 most stable WALs features. We also provided a method for establishing the most stable items for both types of data sets. It might seem obvious that the combination of two methods that each have been used to establish (genetic) relationships between languages would give a result that is at least as good as the best of the two methods individually. However, it is not necessarily clear beforehand *how much better* the results might be. We hope to have shown that an improvement may be expected.

For comparisons and the calibration of the method, the ones with finer genetic distinctions such as the Ethnologue seem to work better than those based on distinctions at the family and genus level only, such as the classifications provided in the WALs database.

Arguably, for the purpose of language classification, be it genetic, areal or otherwise, samples give more reliable and interesting results to the extent that there are more languages in them. In that sense an extensive, worldwide sample based purely on lexical data is much easier to attain than one based on typological material. Therefore, the ASJP project will continue to extend its already large lexical database. Nevertheless, we think that a project that aims at the extension of the WALS database by filling at least part of the around 84% empty cells should be seen as an enterprise of at least equal interest. If language classification were a major motivation, then concentration on the more stable features would be preferable.

### References

- Albatineh, Ahmed N., Magdalena Niewiadomska-Bugaj, and Daniel Mihalko (2006). On similarity indices and correction for chance agreement. *Journal of Classification* 23: 301-313.
- Atkinson, Quentin, Geoff Nichols, David Welch, and Russell Gray (2005). From words to dates: water into wine, mathemagic, or phylogenetic inference? *Transactions of the Philological Society* 103: 193-219.
- Bakker, Dik (2008). LINFER: inferring implications from the WALS database. *STUF* 61-3: 186-198.
- Brown, Cecil H., Eric W. Holman, Søren Wichmann, and Viveka Velupillai (2008). Automated Classification of the World's languages: A description of the method and preliminary results. *Sprachtypologie und Universalienforschung*.
- Dryer, Matthew S. (1992). The Greenbergian word order correlations. *Language* 68: 81-138.
- Dryer, Matthew S. (2005). Genealogical language list. In: Haspelmath et al. (eds.): 584-643.
- Gordon, Raymond G., Jr. (ed.) (2005). *Ethnologue*. 15th Edition. SIL International. <[www.ethnologue.com](http://www.ethnologue.com)>.
- Greenberg, Joseph H. 1978. Diachrony, synchrony and language universals. In: Joseph H. Greenberg, Charles A. Ferguson, and Edith A. Moravcsik (eds.)

- Universals of Human Language, Vol. III: Word Structure*. Stanford: Stanford University Press, 47-82.
- Haspelmath, Martin, Matthew Dryer, David Gil, and Bernard Comrie (eds.) (2005). *The World Atlas of Language Structures*. Oxford: Oxford University Press.
- Holman, Eric W., Christian Schulze, Dietrich Stauffer, and Søren Wichmann (2007). On the relation between structural diversity and geographical distance among languages: Observations and computer simulations. *Linguistic Typology* 11(2): 393-422.
- Holman, Eric, Søren Wichmann, Cecil H. Brown, Viveka Velupillai, André Müller, and Dik Bakker (2008). Explorations in automated language classification. *Folia Linguistica* 42(2): 331-354.
- Huson, Daniel H. (1998). SplitsTree: A program for analyzing and visualizing evolutionary data. *Bioinformatics* 14.10: 68–73.
- Levenshtein, Vladimir I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Cybernetics and Control Theory*, 10(8): 707-710.
- Nakhleh, Luay, Don Ringe, and Tandy Warnow (2005). Perfect phylogenetic networks: A new methodology for reconstructing the evolutionary history of natural languages. *Language* 81: 382-420.
- Nichols, Johanna (1995). Diachronically stable structural features. In: Henning Andersen (ed.) *Historical Linguistics 1993. Selected Papers from the 11th International Conference on Historical Linguistics*, Los Angeles 16-20 August 1993. Amsterdam/Philadelphia: John Benjamins, 337-355.
- Oswalt, Robert L. (1970). The detection of remote linguistic relationships. *Computer Studies in the Humanities and Verbal Behavior* 3: 117-129.
- Rijkhoff, Jan & Dik Bakker (1998). Language sampling. *Linguistic Typology* 2(3): 263–314.
- Sapir, Edward (1970 [1921]). *Language: An Introduction to the Study of Speech*. London: Rupert Hart-Davis.
- Serva, Maurizio and Filippo Petroni (2008). Indo-European languages tree by Levenshtein distance. *EuroPhysics Letters* 81: 68005.

Swadesh, Morris (1955). Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics* 21: 121-137.

Wichmann, Søren and Eric W. Holman. (n. d.). Assessing temporal stability for linguistic typological features. Manuscript under review. Prepublication version:

<http://email.eva.mpg.de/~wichmann/WichmannHolmanIniSubmit.pdf>.

---

<sup>1</sup> Bakker is responsible for write-up of this paper. C. Brown, Holman, and Wichmann have contributed with research design and analysis as well as data, and Velupillai, Müller, P. Brown, Egorov, Mailhammer, and Grant have contributed with data and discussion.

<sup>2</sup> We excluded from the start three variables for which a very restricted number of languages have a value.

<sup>3</sup> These features may be either of an areal or genetic nature. In the case of the latter, the time depth is taken to be anywhere between 4000 and 6000 years.

<sup>4</sup> In fact, it may even become negative in case the value concerned is more common between unrelated language pairs than related ones.

<sup>5</sup> Arguably, the same holds for measuring the distances between linguistic features. As stated above, we now measure in terms of equality only. However, some feature pairs may be much closer to each other than others in terms of diachronic change. Currently, we do not have enough information at our disposal for most of the WALS features in order to weigh the distances between the individual values more accurately.

<sup>6</sup> This figure is analogous to figure 5 in Holman et al. (2008).

<sup>7</sup> Experiments have shown that the results for 1,000 randomizations do not substantially diverge from larger numbers.